DOCUMENT RESUME

ED 200 631 TM 810 217

AUTHOR Diamond, Esther E.
TITLE Item Bias Issues: Background, Problems, and Where We Are Today.
PUB DATE Apr 81
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; Psychometrics; *Test Bias; Testing Problems; *Test Items; Test Validity
IDENTIFIERS Content Validity; Criterion Related Validity; Social Justice; Test Publishers

ABSTRACT

As test standards and research literature in general indicate, definitions of test bias and item bias vary considerably, as do the results of existing methods of identifying biased items. The situation is further complicated by issues of content, context, construct, and criterion. In achievement tests, for example, content validity may impose demands that necessitate inclusion of some items identified as biased or marginally biased. Or--a question raised by a number of researchers--after biased items have been eliminated or, within the limitations of content requirements, have been balanced with items biased in the other direction, does the test measure the same construct for each group under consideration? Other questions involve the assessment of test fairness in the presence of external criteria and in the absence of such criteria, as well as the determination of what constitutes an adequate criterion measure. The state of the art with reference to these questions and to the issues of social justice associated with test bias is reviewed, as well as procedures employed by various publishers of achievement and aptitude tests to detect item bias and to minimize bias in the final tests. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200631

Item Bias Issues: Background, Problems, and Where We Are Today[1]

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E. E. Diamond

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Esther E. Diamond

Science Research Associates

[1]Paper presented at symposium, Investigating Test Bias from Item Preparation Through Standardization and Beyond, at annual meeting of the American Educational Research Association, Los Angeles, April 1981. The views expressed here are the author's and do not necessarily represent those of Science Research Associates.

TM 810 217

Running Head: Item Bias Issues

Item Bias Issues: Background, Problems, and Where We Are Today

Esther E. Diamond

Science Research Associates

Abstract

As the APA/AERA/NCME test standards and the research literature in general indicate, definitions of test bias and item bias vary considerably, as do the results of existing methods of identifying biased items. The situation is further complicated by issues of content, context, construct, and criterion. In achievement tests, for example, content validity may impose demands that necessitate inclusion of some items identified as biased or marginally biased. Or--a question raised by a number of researchers--after biased items have been eliminated or, within the limitations of content requirements, have been balanced with items biased in the other direction, does the test measure the same construct for each group under consideration? Other questions involve the assessment of test fairness in the presence of external criteria and in the absence of such criteria, as well as the determination of what constitutes an adequate criterion measure. This presentation will briefly review the state of the art with reference to these questions and to the issues of social justice associated with test bias. It will also review the procedures employed by various publishers of achievement and aptitude tests to detect item bias and to minimize bias in the final tests.

Item Bias Issues: Background, Problems, and Where We Are Today

Esther E. Diamond

Science Research Associates

Bias in testing and related issues such as bias in selection and the fair use of tests have become major concerns for the educational and psychological measurement profession over the past decade. Papers on the issues and on purported ways of dealing with them abound in the professional literature, conferences, and conventions. The courts, too, have been kept busy trying to resolve problems of test and selection bias at the social issues level.

This paper, and the symposium of which it is a part, deals mainly with bias in achievement tests. Bias in selection has been addressed extensively in the literature, especially in a special issue of the Journal of Educational Measurement, On Bias in Selection (1976).

Definitions of Bias

Definitions of bias are as profuse as the suggested methods for correcting it. Standards for Educational and Psychological Tests (American Psychological Association, 1974) alerts test users to the existence of the variety of definitions for bias and fairness and points out that whether a given procedure is fair may depend on the definition accepted.

Early on, Cleary (1968) suggested that a test is biased if scores for subgroups are consistently predicted too high or too low. Green (1975) defined a biased test as one that produces results systematically unfair to some group, measuring variables for the group that are at least partly distinct from those it measures for others in the population. Unfairness,

Green maintained, is not in the nature of the test but in the way in which scores are used.

Other definitions deal with bias at the item level. Jensen (1980) views test bias as a statistical concept, one that can be appropriately evaluated only by means of empirical analysis. Most claims of test bias, he avers, are not based on such analysis but are, rather, expressions of particular social philosophies. He considers two basic criteria for establishing test bias--external or predictive validity, and internal or construct validity.

Scheuneman (1975) considered an item unbiased if, for persons with the same ability in the area being measured, the probability of a correct response is the same regardless of the population group of which the individual is a member. That is, an item is unbiased if, for all individuals with the same score on a homogeneous subtest containing the item, the proportion that answers the item correctly is the same for each population group taking the test.

Linn, Levine, Hastings, and Wardrop (1980) were critical of definitions of item bias that depend on differences in proportion correct for two groups. Proportion correct, they maintained, can be confounded with other item characteristics such as discriminating power of the items.

Shepard, Camilli, and Averill (1980) defined an item as biased if two individuals with equal ability but from different groups do not have the same probability of success on the item. In an achievement test, Shepard claimed, bias would be suspected if children in one group

consistently receive lower scores than would be expected from observed classroom performance. Cole (1978) commented that in achievement tests, as in selection, the same underlying definition has dominated: An item is thought to be biased if it does not measure the same things in different groups.

There are still other definitions of test and item bias, most of them variations of those offered here, depending on the particular methodology used to identify the bias.

Psychometric Item-Bias Detection Methods

Judgmental methods of detecting bias--editorial reviews and reviews by panels representative of the groups whom bias impacts most, at successive steps in the item-specification and item-writing stages--are important first steps and will be discussed in greater detail later in this paper. But, as Shepard (1980) pointed out, both logical and empirical evidence of bias are required to verify that what we think is being measured for all groups is indeed being measured. "When faulty conceptualizations of a trait produce different implicit inferences for different groups, statistical techniques reveal the error missed by logical analysis" (p.20).

Like definitions of bias, a number of methods of detecting bias through psychometric techniques are described in the literature. They include converting p-values to normal deviates and plotting each item in a bivariate plot; determining point biserials separately for each comparison group and comparing the distributions; using analysis of variance to obtain

an estimate of between-group variation for each item and eliminating those items for which the estimate is high, the goal being to include items that show within-group variation consistently; using various latent trait or item-characteristic-curve models such as the Rasch one-parameter model and the three-parameter item characteristic curve; and using chi-square to compare obtained and expected p-values.

Linn et al. (1980) pointed out the difficulty in identifying systematic differences between biased and unbiased items, or between items biased in different locations:

> For example, items asking the meaning of a word in context sometimes appeared to be biased in one direction and sometimes in the other. Thus, no generalized principles that would be useful in avoiding items that tend to show bias can be stated for guiding the future construction of tests of reading comprehension. Instead, only a post-hoc analysis procedure...can be offered (pp. 30-31).

Berk (1980) emphasized that the content or behavior a test measures is an integral part of all score inferences, and an item is the most fundamental level of content analysis and the foundation for score inferences. Item bias studies, consequently, are necessary for all tests, but they are not sufficient for all test score inferences and uses--for example, predicting future performance. Furthermore, Berk maintained, the results of bias studies at the subtest and total test levels do not preclude the presence of bias at the item level, although elimination of item bias may decrease the likelihood of test bias.

To examine bias at the item distribution level, Medley and Quirk (1974) examined differences between black and white candidates' performance on the common examinations of the National Teacher Examinations. They

constructed experimental forms and compared performance on items reflecting black culture, those reflecting modern culture, and those that could be defined as traditional. Differences in performance on one test that was made up of equal numbers of black and modern-culture items and on another test consisting solely of traditional items were significant for 13 of the 14 pairs of groups tested. Significant differences were also found in favor of blacks on the black-culture items and in favor of whites on the modern-culture items.

Echternacht (1974) compared the distributions of transformed p-value differences for independent pairs of groups with a hypothetical normal distribution; the obtained mean and variance of the differences were used as parameters. Echternacht considered the test biased if points on the actual distribution fell outside the bands around a hypothetical line whose width was determined by sample size and significance level.

In Angoff's delta plot procedure (Angoff & Sharon, 1974), item difficulties were transformed into deltas--normal deviates with a mean of 13 and a standard deviation of 4. Items were considered biased if deltas differed substantially for two comparison groups, based on their perpendicular distance from the major axis of the elliptical delta plot. Angoff (1975) described several studies, including his own, in which bivariate plots of transformed p-values were examined for item x group interaction. Angoff considered the possibility of building a test on the basis of a common core of items "broadly relevant to the educational objectives of society generally and the individuals for whom it is intended" (p. 26) as well as items specific to the curriculum of each of the component groups, but not the group as a whole. With such balance, Angoff speculated, no one group would have an advantage across the total test.

In one of the studies described by Angoff (1975), involving black and white groups, item x group interaction for inter-race scatterplots decreased when the groups were matched on an external variable. This suggests the possibility of matching groups on socioeconomic status, although Angoff warns that the designations for SES levels might not have exactly the same meanings for blacks as for whites.

Scheuneman (1980) experimented with three different methods of item bias analysis to see how consistent the results would be across two administrations of a test. An item was defined as biased if it was high on one index and at least moderate on the second. For all three procedures--delta, Rasch, and her own chi-square procedure--items identified as biased on both occasions agreed more often than would be expected by chance. Nevertheless, Scheuneman found the degree of agreement disappointing. But using a combination of the three procedures--particularly chi-square with either of the other two--worked better than any one procedure by itself.

Responding to a criticism of her chi-square procedure, Scheuneman (1981) redesignated it as C2, pointing out that it is a statistic "similar to chi-square" (p. 63) and reiterating an earlier statement that the obtained distribution of values may not always approximate the appropriate chi-square distribution, particularly if sample sizes for the groups being compared are quite different or the cell frequencies very large. She pointed out further that, rather than forming a biased/unbiased dichotomy, "bias is more likely a matter of degree with some items less biased than others" (pp. 64-65) and therefore one might question the necessity of C2 being chi-square distributed.

Models based on latent trait theory--such as the Rasch model, which considers only one parameter, that of difficulty, and assumes that all items in a test or subtest have equal discriminating power, and the three-parameter model, which considers difficulty, discrimination, and guessing, or the probability of an individual with extremely low ability getting the item right--are probably the most widely discussed and employed models for dealing with item bias today. (For a comprehensive discussion of latent trait models, see particularly the 1977 special issue of the Journal of Educational Measurement, Applications of Latent Trait Models, as well as Hambleton, 1980.) From a theoretical point of view, the three-parameter model (Lord, 1977) appears to be the most inclusive and the preferred model. In practice, however, it is generally conceded that the program is the most difficult one to implement; it cannot be run economically with a very limited number of items or with relatively small samples of persons tested. While the advances in latent trait analyses are hopeful, there still appear to be a number of unresolved issues, including the assumption of unidimensionality of ability, and what happens when the test data do not fit the particular model.

Content, Context, and Construct

Questions of item content, context, and construct further complicate the problem of identifying--and correcting--item bias in achievement tests. Ebel (1975), for example, viewed content validity as the primary characteristic to be sought in achievement testing. A test of educational achievement, he maintained, is not biased if it consists of tasks that sample representatively a clearly defined domain of tasks and if it yields reliable scores. Cole (1978), however, pointed out that measurement experts and test users often forget that content validity evidence supports

only content-referenced interpretations and not inferences beyond the particular content domain. We are often unable to define all the crucial limiting aspects of the content domain--for example, the format of the item, the language required, and the level of reading and vocabulary. "We label achievement tests with construct-type labels such as math concepts, arithmetic reasoning, and reading comprehension" (p. 6), rarely limiting our interpretations to the cultural content domain of the test.

Cole cited as an example the Math Concepts Test of the Stanford Achievement Test, which suggests inferences beyond the actual content domain, offering a construct-like interpretation. Such interpretations, Cole pointed out, are the most common types of interpretations made of achievement test scores. "When considering bias in achievement tests, we are necessarily concerned about the extent to which the test in fact measures the same construct for different groups (p. 13). For example, if a test measures a construct-like concept for one group and we apply that interpretation to a second group, for the latter the test may be measuring something different. That is, a test might measure a mathematics concept for the first group; for the second group, however, it might be measuring vocabulary, in which the mathematics concept was presented.

Dwyer (1976) found that in tests of verbal ability, all other things being equal, males obtained higher scores when the material was set in the context of business, science, practical affairs, mechanical principles, or mathematics. Females scored higher if the material was drawn from the arts, the humanities, or understanding human relations. There was no conclusive

evidence as to whether these differences were the result of familiarity with the context, or motivational considerations associated with the context, or some combination of the two.

Diamond (1979) pointed out that sex bias in item content may arise from the fact that the item pool in an area traditionally considered one in which males are dominant tends to be based on male experiences or to use male referents. Twenty years ago Coffman (1961) drew similar conclusions from his study of the differences in the performance of men and women on a verbal aptitude test. He found that people-oriented words favored women and things-oriented words favored men and concluded that "[T]he observed differences must reflect different experiential backgrounds of men and women in our culture. Similarly, there must be differences in the experiential backgrounds of people classified in other ways." If the differences are relevant for the performance one wishes to predict, Coffman advised, then items showing differences should remain in the test. If, however, the differences are not relevant, they "should be removed from the item pool or controlled at the point of assembly to insure optimum weighting in the test" (p. 9).

Tittle (1978) cited a 1975 study by McCarthy that showed the effects of item context on performance in mathematics at the high school level. While the basic mathematical processes remained the same, when item context was manipulated experimentally to describe materials or settings more familiar to males or to females, sex differences in performance resulted.

Shepard (1980) suggested that freedom from bias in test items is best conceptualized as equal construct validity across groups. And while in

the past achievement tests have ordinarily needed only content validation rather than construct validation, measurement experts are now suggesting that even criterion-referenced achievement tests should show evidence of construct validity. Empirical support, Shepard maintained, is needed for the logical arguments about what the test measures. In an earlier paper, Shepard et al. (1980) pointed out that bias cannot be identified in isolated test items. Items designed to measure the same construct must be studied together; bias is discovered when an item does not fit the pattern established by other items in the set.

Tenopyr (1977) described attempts to distinguish between the content and construct aspects of validity as resulting in "the most confusion among both practitioners and theoreticians" (p. 47). Content-oriented test development must always take constructs into consideration. Messick (1980) agreed, viewing the attempt to define content validity separately from construct validity as producing "a disfunctional strain to avoid constructs, as if shunning them in test development somehow lessens the impact of response processes in test performance" (p. 1018).

The Criterion Question

A major limitation of bias detection approaches to date, as Shepard et al. (1980), among others, have suggested, is that they are based on a criterion internal to the test in question. "They cannot escape the circularity inherent in using total score on the test or the average item to identify individuals of equal ability and hence specify the standard of unbiasedness" (p. 6). Nor is an external criterion the answer if it is contaminated by the same type of bias that affects the items. A number of other writers (see, for example, Wild & Dwyer, 1978) have pointed out the fallacies in assuming that the external criterion,

especially in models for selection, is a valid, reliable, and unbiased measure.

Linn/(1980) et al. expressed a similar point of view: When the total test score is used as the criterion, prediction cannot be the purpose, since the total score can generally be perfectly predicted by the set of items on which it is based. Besides, when items are discarded as a result of item analysis, the total score will change.

When there is an outside criterion, Angoff (1980) commented, "we have at our disposal an agreed-upon relationship...external to the realm of testing, against which to evaluate the test" (p. 1). Examining test bias without an outside criterion, however, is much more elusive.

There are also instances, Angoff pointed out, "when prediction is irrelevant to the conceptual test of bias" (p. 2), as in the case of achievement testing, unless we use course grades or grade point average as criteria. Evaluation of bias without an external criterion, Angoff maintained, must come from an examination of the internal components of the test relative to one another or the examination of test data "in the context of other empirical data and theory independently developed" (p. 2).

The Question of Social Justice

Let us return for a moment to Jensen's statement about bias in testing (1980) cited earlier in this paper. Jensen, you will remember, views test bias as a statistical concept, one that can be appropriately evaluated only (emphasis ours) by means of empirical analysis, and he maintains that most claims of test bias are not based on such analysis

but are, instead, expressions of particular social philosophies. His is not a unique view; it is shared by many--professional test developers and researchers as well as test "consumers." Others express what might be called a social justice point of view. For example, Messick (1980) reiterated an earlier position that not only should tests be evaluated in terms of their measurement properties, but testing applications should be evaluated in terms of potential social consequences. Messick pinpointed two critical questions: (1) Is the test good as a measure of the characteristics it is interpreted to assess? (2) Should the test be used for the proposed purpose in the proposed way? The first question is concerned with the psychometric properties of the test, especially construct validity. The second question is an ethical one, and to answer it the proposed use of the test must be justified in terms of social values.

These questions, Messick held, apply not just to matters of selection or personality assessment but to all psychological and educational measurement, including construct-based ability tests and content-sampled achievement tests. Such measurement

> entails value judgments--at all levels of test construction, analysis, interpretation, and use...Values thus appear to be as pervasive and critical for psychological and educational measurement as is testing's acknowledged touchstone, validity (p. 1013).

Similarly, Angoff (1980) viewed item fairness and test fairness as concepts that must be evaluated in terms of purpose of the instrument and its intended use, and not just on the basis of statistics alone. The presence or absence of bias, he maintained, is a matter for human judgment, "for which item analysis is only a useful tool" (p. 28).

-Katz
Lockheed/(1974), Dwyer (1976), and Shepard (1980), among numerous others, have taken the position that whether or not performance is affected differentially by content representing males and females or ethnic groups unequally and stereotypically is beside the point. Such effects are difficult to measure; too many uncontrollable variables get in the way. We do not know, for example, to what extent such differential content affects self-concept, or motivation, or role perceptions. As a matter of simple human justice, however, any aspect of a test that is offensive to members of a group should routinely be eliminated from the tests. That this is not always the practice, despite growing emphasis on the importance of doing so, is illustrated in part in a study by Zaref and Williams (1980). They examined six leading intelligence tests and found overwhelming sexual and racial imbalance in the content, with frequent racial stereotyping and 100 percent sex-role stereotyping for some of the tests.

Perhaps it would be appropriate to say something here about the various attempts over the years to build "culture-free," "culture-fair," and "culture-specific" tests--usually tests of intelligence but sometimes suggested as replacements for standardized achievement tests. Nonverbal culture-fair tests, as Ornstein (1976) pointed out, generally fail to reflect the full range of a child's mental abilities. Moreover, the child who has trouble with verbal tasks generally has trouble with such perceptual tasks as classification, selection, and arrangement. As for the culture-specific Black Intelligence Test of Cultural Homogeneity (BITCH), it has been criticized as measuring a very limited amount of special information useful for functioning in the ghetto. Other abilities, important for <u>all</u> children if they are to succeed in school, are not

dealt with.

What Test Publishers Are Doing About Bias

Following the enactment and implementation of legislation affecting civil rights and sex equity in the late '60s and early '70s, professional organizations and publishers of educational materials and tests began issuing "fairness" guidelines. As a survey of sex bias in achievement testing completed in 1980 by the Commission on Sex Bias in Measurement of the Association for Measurement and Evaluation in Guidance (AMEG) indicated, however, most of the guidelines dealt with "facial" bias, and most publishers were monitoring achievement tests at various stages of editorial item development only, for both ethnic and sex bias (Diamond, 1980). Most publishers were paying attention to frequency and role of male and female and ethnic characters in the test, and to the use of nonsexist language. Pretest data for only some of the tests covered in the survey were analyzed separately for sex (or ethnic) differences, most of these were tests developed after 1970, and a number of publishers indicated that with the development of new instruments or with revisions of earlier isntruments, such analyses would be conducted. Pretest analyses, where they were conducted, involved use of a number of the psychometric techniques described in this paper--ranging from a simple comparison of p values to the use of Rasch analyses. Where biased items were identified, publishers generally eliminated them if content considerations permitted doing so, or they compensated with items biased in the other direction.

Carlton and Marco (1980) reported that Educational Testing Service has used refinements in delta plot methodology and has also used random samples

to establish a baseline for comparisons among groups. In addition, it has conducted an investigation of error response patterns in the National Teacher Examinations, has studied differential item performance on the Scholastic Aptitude Test on a routine basis, has instituted a test sensitivity (judgmental) review procedure, and has employed item response theory methods and log-linear analysis.

Coffman (1980) described the methods of bias analysis used by Riverside Publishing Company--an eclectic approach representing both subjective judgments and statistical analyses based on several models for bias-detection. Judgmental reviews of pretest forms by potential users turned up observations not previously noted by item writers and test constructors. As for statistical analyses, no one method proved to be clearly superior. Results with various tests, based on the arcsin transformation method, ANOVA, and the three-parameter item-characteristic-curve model, were not comparable. Further exploratory studies are being conducted.

Lenke (1980) reported on the various techniques currently being used by The Psychological Corporation to detect bias in tests currently or recently in some stage of new development, including revision and re-standardization. Tests published prior to 1976, Lenke observed, were not routinely subjected to item bias procedures during their development, but the more recent the test, the more well defined the procedures--until they are an integral part of the test development process today.

Raju (1980) and Green (1980) also reported increasing attention to item bias detection procedures currently employed by Science Research Associates

and California Test Bureau/McGraw Hill, respectively. Again, a combination of judgmental and statistical procedures are employed.

In the development of SRA's Achievement Series 1 and 2, for example, the effort to minimize bias began with detailed specifications for item writers, designed to make them sensitive to all the possibilities for editorial bias. Use of unbiased language and balance in ethnic and sex work roles and life styles, as well as in numerical representation, were stressed. Women and minorities were positively portrayed in a wide range of traditional and nontraditional roles. Item writers were cautioned against tokenism and the use of negative stereotypes.

At various stages during the item development process, items were reviewed by content and testing specialists and by knowledgeable representatives of minority groups and women. Acceptable items were then pretested with both majority and minority groups. Those items that survived the first statistical screening were then subjected to four separate analyses--female vs. male, black vs. white, Hispanic vs. white, and American Indian vs. white, using transformed item p-values (delta values), as described by Raju (1980). Still further analyses will be described by Bode (1980) as part of this symposium.

For the California Achievement Tests Forms C and D, CTB/McGraw Hill went through similar screening processes through the item development stage and the initial screening of tryout data. In the second screening of tryout data, data for black students were examined separately and compared with data for students identified as Other. Content validity was the primary consideration, and so some items were included even though they

appeared to be biased or otherwise unsatisfactory on the basis of the tryout data. If possible, they were revised. The item statistics used as indicators of bias were point biserial correlations between items and category-objective scores, and grade-to-grade growth as shown by item difficulties.

Analysis of the most recent form of the California Test of Basic Skills illustrates the increasing level of sophistication that has characterized CTB's use of item-bias methodology from 1968 to date. Data from Form U, based on the three-parameter latent trait model, were examined for differential item locations for different groups.

With all the advances made by researchers and test publishers, there is still a long way to go. The problems of item bias may require much more than the most sophisticated of statistical procedures for a solution. In the process of seeking the solution, however, we will undoubtedly learn much more about the nature of the beast, and even--just possibly--how to prevent it from manifesting itself in the first place.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: Author, 1974.

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, Md., December 1975.

Angoff, W.H. The use of difficulty and discrimination indices in the identification of biased test items. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Angoff, W.H., & Sharon, A.T. The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 1974, 34, 807-816.

Applications of latent trait models. Special issue, Journal of Educational Measurement, 1977, 14, 73-196.

Berk, R.A. Introduction to test item bias methodology: The state of the art. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Bode, R.K. Comparison of pretest and re analysis results of an item bias study. Paper presented at the nnual meeting of the American Educational Research Association, Los Angeles, April 1981.

Carlton, S.T., & Marco, G.L. Bias methods used by test publishers: Educational Testing Service. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Cleary, T.A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Coffman, W.E. Sex differences in responses to items in an aptitude test. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February 1961.

Coffman, W.E. Bias methods used by test publishers: Riverside Publishing Co., Houghton-Mifflin. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Cole, N.S. Approaches to examining test bias in achievement test items. Paper presented at the annual meeting of the American Personnel and Guidance Association, Washington, D.C., March 1978.

Diamond, E.E. Sex equity and measurement practices. In Lennon, R.T. (Ed.), New Directions for Testing and Measurement, 3: Impactive Changes on Measurement. San Francisco: Jossey-Bass, 1979.

Diamond, E.E. The AMEG Commission report on sex bias in achievement testing. Measurement and Evaluation in Guidance, 1980, 13, 135-147.

Dwyer, C.A. Test content in mathematics and science: The consideration of sex. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Ebel, R.L. Educational Tests: Valid? Biased? Useful? Phi Delta Kappan, 1975, 57, 83-89.

Echternacht, G.A. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Green, D.R. What does it mean to say a test is biased? Education and Urban Society, 1975, 8, 33-52.

Green, D.R. Bias methods used by test publishers: California Test Bureau/ McGraw Hill. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Hambleton, R.K. Latent trait models and their applications. In Traub, R. (Ed.), New Directions for Testing and Measurement, 4: Methodological Developments. San Francisco: Jossey-Bass, 1979.

Jensen, A.R. Bias in mental testing. New York: The Free Press, 1980.

Lenke, J.M. Bias methods used by test publishers: The Psychological Corporation. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. An investigation of item bias in a test of reading comprehension. Technical Report No. 163. Champaign, Ill.: University of Illinois at Urbana-Champaign, March 1980.

Lockheed-Katz, M. Sex bias in educational testing: A sociologist's perspective. Research Memorandum. Princeton, N. J.: Educational Testing Service, 1974.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Medley, D.M., & Quirk, T.J. The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examinations. Journal of Educational Measurement, 1974, 11, 235-245.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

On bias in selection. Special issue, Journal of Educational Measurement, 1976, 13, 1-100.

Ornstein, A. IQ tests and the culture issue. Phi Delta Kappan, 1976, 57, 403-404.

Raju, N.S. Bias methods used by test publishers: Science Research Associates. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Scheuneman, J. Consistency across administrations of certain indices of bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Scheuneman, J. A response to Baker's criticism. Journal of Educational Measurement, 1981, 18, 63-66.

Shepard, L. Definitions of test bias. Paper presented at the Johns Hopkins University Symposium on Educational Research--Test Item Bias Methodology: The State of the Art, Washington, D.C., November 1980.

Shepard, L., Camilli, G., & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Tenopyr, M.L. Content-construct confusion. Personnel Psychology, 1977, 30, 47-54.

Tittle, C.K. Sex bias in achievement testing: Policies, issues, and implications for action. Paper presented at the annual meeting of the American Personnel and Guidance Association, Washington, D.C., March 1978.

Wild, C.L., & Dwyer, C.A. Sex bias in selection. Research memorandum. Princeton, N.J.: Educational Testing Service, February 1978.

Zoref, L., & Williams, P. A look at content bias in IQ tests. Journal of Educational Measurement, 1980, 17, 313-322.